ANALYSIS OF IMAGES SOCIAL NETWORKS AND TEXTS

Bigram Anchor Words Topic Model

Ashuha Arseniy, Loukachevitch Natalia

Moscow Institute of Physics and Technology
Research Computing Center of Lomonosov Moscow State University

ars.ashuha@gmail.com louk_nat@mail.ru

April 29, 2016

Motivation

- Nowadays we have a lot of data, but usually **it is unlabled**
- We wont to extract structure from document collection **unsupervised**

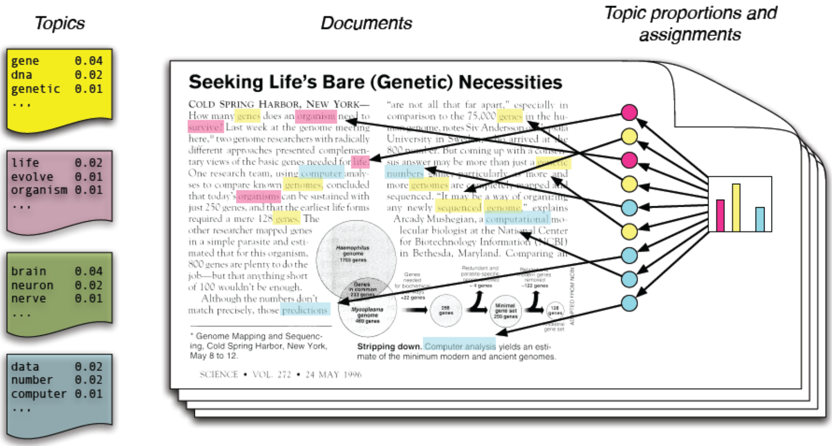
Haw can we get this goal?

- Topic modeling is a powerful tool for document collection analysis
- Unformally, topic is a semantically related set of words

  sample: *geom rna fast dna sequence alignment nucleotides*

More formal

- Topic is a discrete distribution over words $p(w|t) = p(word|topic)$
- Document is a discrete distribution over topics $p(t|d) = p(topic|doc)$
- We want to find $p(w|t)$, $p(t|d)$ given $p(word|doc)$
- Usually we solve this problem as a matrix decomposition

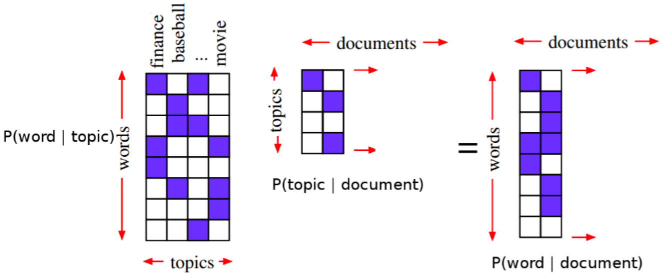

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Probabilistic model

$$p(word|doc) = \sum_{topic} p(word|topic)p(topic|doc)$$

- The order of words in document is not matter (bag of words)
- Topic is not depends on doc ($p(word|doc, topic) = p(word|topic)$)

Represent it as matrix decomposition and solve this problem by MLE

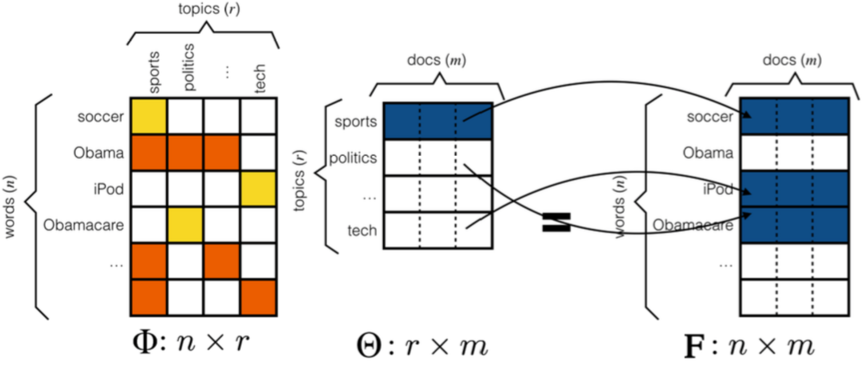


Regularization

- LDA – topics and documents generated from Dirichlet distribution
- BigARTM – generalize LDA, many regularizes

- $+$ Good matrix approximation
- $+$ A lot of implementations
- $+$ There exist modification to take into account bigrams
- $-$ Solution is really depends on initial approximation
- $-$ Poor model of documents
- $-$ Difficult to parallelize
- $-$ Computational difficult
- $-$ Control coefficient of regularizations is really hard task

Let's assume for each topic T there exist word w that $p(w|t) \neq 0 \; if \;\; t = T$



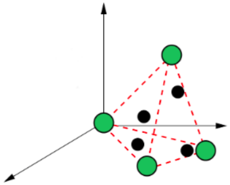
Therefore F is a just a linear composition constructed Θ rows, anchor rows.

1. How can we found rows in F which corresponds to anchor words?
2. How can we reconstruct topic model (Φ, Θ) given anchor words?

- Matrix F too noisy, let's use FF^t
- Size of FF^t is $Words \times Words$ therefor reduce dimension

$$FF^t_{words \ \times \ words} = H_{words \times k}$$

- Find almost convex hull in rows of H matrix $\{H_{anchor_1}, ..., H_{anchor_n}\}$



- Solve **undependent** convex optimization problems: find c_i for each t

$$H_t \approx \sum_{i=1,\dots,T} c_{ti} H_{anchor_i}, \quad c_{it} \geq 0, \quad \sum_i c_{it} = 1, c_i = p(topic|word)$$

- Use Bayes rule to reconstruct $\Phi = (p(word|topic))_{W \times T}$

+ No initial approximation
+ Very well parallelize out of box
– Need to tune parameters
– Can't take into account bigrams
– Worst matrix decomposition

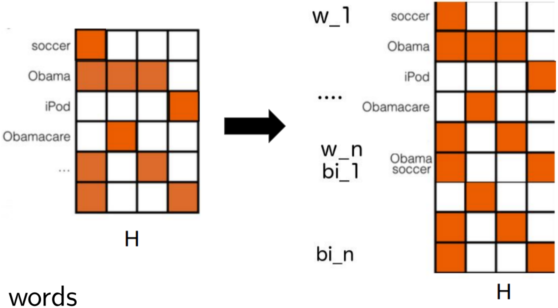
Our goal was propose modification witch **can take into account bigrams**.

Why it is important?

- adding new information about word order in model
- better and lighter interpretability
- simple solution – does not work

There is one simple way:

1. precomputed bigrams
2. we assume that vector for bigram $w_i w_j = H_{w_i} + H_{w_j}$
3. add vectors corresponds bigrams to set of points H (finding anchors)



4. Find anchor words
5. Recover topic mode
6. Make some PLSA steps

Bigrams can be anchor words

Interpretations good latent space in matrix H

Old anchors

- loss
- cluster
- mixtur
- synaps
- theorem
- speech
- entropi
- filter
- competit
- gain
- markov
- identif
- algorithm

Our anchors

- mixtur
- boltzmann_machin
- likelihood
- markov_chain
- action
- vector_quantiz
- network
- robot_arm
- loss
- tangent_distanc
- classifi
- reinforc_learn
- speech

Metrics:

- **Perplexity** is a mean $exp(-mean\ likelihood)$
- **Coherence** is a mean Pointwise Mutual Information
- **Unique of kernels** is a mean Jaccard distance between most probable words in topic

Collection	**Banks Articles**			**20 Newsgroups**			**NIPS**		
Metric	P_{test}	PMI	U	P_{test}	PMI	U	P_{test}	PMI	U
PL	2116	0.60	0.40	2155	0.31	0.40	1635	0.21	0.32
AW	2330	0.63	0.53	2268	0.38	0.41	1505	0.41	0.38
BiAW	2248	0.79	0.60	2183	0.68	0.54	1500	0.50	0.41
AW+PL	2052	0.78	0.58	2053	0.54	0.55	1434	0.52	0.46
BiAW+PL	**1848**	**0.87**	**0.63**	**2027**	**0.78**	**0.64**	**1413**	**0.58**	**0.49**

- Arora, S., Ge, R., Moitra, A.: Learning topic models - going beyond svd. In: Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, IEEE (2012) 1–10
- Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. arXiv preprint arXiv:1212.4777 (2012)
- B, Dobrov, N, Loukachevitch: Forming the base of terminological phrases in the texts of the subject area (2003) 201–210